Fusion of children's speech and 2D gestures when conversing with 3D characters

Jean-Claude Martin[a,*], Stéphanie Buisine[a], Guillaume Pitel[a], Niels Ole Bernsen[b]

[a]*Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI-CNRS), BP 133, 91403 Orsay Cedex, France*
[b]*Natural Interactive Systems Lab, Campusvej 55, DK 5230 Odense M, Denmark*

Abstract

Most existing multi-modal prototypes enabling users to combine 2D gestures and speech input are task-oriented. They help adult users solve particular information tasks often in 2D standard Graphical User Interfaces. This paper describes the NICE Andersen system, which aims at demonstrating multi-modal conversation between humans and embodied historical and literary characters. The target users are 10–18 years old children and teenagers. We discuss issues in 2D gesture recognition and interpretation as well as temporal and semantic dimensions of input fusion, ranging from systems and component design through technical evaluation and user evaluation with two different user groups. We observed that recognition and understanding of spoken deictics were quite robust and that spoken deictics were always used in multi-modal input. We identified the causes of the most frequent failures of input fusion and suggest possible improvements for removing these errors. The concluding discussion summarises the knowledge provided by the NICE Andersen system on how children gesture and combine their 2D gestures with speech when conversing with a 3D character, and looks at some of the challenges facing theoretical solutions aimed at supporting unconstrained speech/2D gesture fusion.

Keywords: Multi-modal interface; Design and evaluation; 2D gestures; Children; Conversational agent

1. Introduction

Since Bolt's seminal Put-that-there paper which heralded multi-modal interaction [1], several system prototypes have been developed that enable users to interact through combined speech-gesture input. It is widely recognised today that this form of multi-modal input might constitute a highly natural and intuitive multi-modal "compound" which all or most humans use for many different communicative purposes. However, most of those prototypes are task-oriented, i.e., they help the user solve particular information tasks in more or less standard Graphical User Interface (GUI) environments. Moreover, the target user group tends to be adults rather than children. This dominant paradigm of GUI-based task-oriented information systems for adults only addresses a fraction of the potentially relevant domains of application for using combined speech and gesture. Outside the paradigm we find, for instance, systems for children, non-task-oriented systems, systems for edutainment and entertainment, and systems for making-friends conversation

*Corresponding author. Tel.: +33 6 84 21 62 05.
E-mail addresses: martin@limsi.fr (J.-C. Martin), buisine@limsi.fr (S. Buisine), pitel@limsi.fr (G. Pitel), nob@nis.sdu.dk (N. Ole Bernsen).

with 3D embodied characters. The challenges to combined speech–gesture input technologies posed by systems like those, including systems which include all of the extra-paradigm properties mentioned, have not been addressed yet to any substantial extent. No existing theory can provide reliable predictions for questions, such as: how do children combine speech and gesture? Would they avoid using combined speech and gesture if they can convey their communicative intention in a single modality? Is their behaviour dependent upon whether they use their mother tongue or a second language? To what extent would the system have to check for semantic consistency between their speech and the perceptual features of the object(s) they gestured at? How to manage temporal relations between speech input, gesture input and multimodal output? How do we evaluate the quality of such systems? What do the target users think of them?

This paper addresses the questions and issues mentioned above in the context of system prototype development and evaluation. We discuss issues in semantic input fusion of speech and 2D gesture, ranging from systems and component design through technical evaluation and user evaluation to taking a look at the future challenges which the work reported has uncovered in a very concrete manner. The work reported was carried out in the EU project NICE on Natural Interactive Communication for Edutainment 2002–2005 (www.nicepro-ject.com). The NICE project has developed two prototypes of each of two related systems, one for conversation with fairytale author Hans Christian Andersen and one for playful computer game–style interaction with some of his fairytale characters in a fairytale world. As we shall focus on the Andersen system below, we would like to point out here that both systems are the results of extensive European collaboration, as follows. For both systems, Swedish computer games company Liquid Media did the graphics rendering; Scansoft, Germany, trained the speech recognisers with children's speech; and LIMSI-CNRS, France, did the 2D gesture components and the input fusion. What makes the two systems different is that the Andersen system's natural language understanding, conversation management, and response generation components were built by NISLab, Denmark, whereas the corresponding components for the fairytale world system were built by Telia-Sonera, Sweden.

1.1. Goals of the NICE Andersen project

The main goal of Andersen system development is to demonstrate natural human–system interaction for edutainment by developing natural, fun and experientially rich communication between humans and embodied historical and literary characters. The target users are 10–18 years old children and teenagers. The primary use setting for the system is in museums and other public locations. Here, users from many different countries are expected to have English conversation with Andersen for an average duration of, say, 5–20 min. The main goal mentioned above subsumes a number of sub-goals, none of which had been achieved, and some of which had barely been addressed, at the start of NICE, i.e. to:

- demonstrate domain-oriented spoken conversation as opposed to task-oriented spoken dialogue, the difference being that, in domain-oriented systems there are no tasks to be performed through user-system interaction. Rather, the user and the system can have free-style, fully mixed-initiative conversation about any topic in one or several semi-open domains of knowledge and discourse;
- investigate the challenges involved in combining domain-oriented spoken conversation input with 2D gesture input;
- investigate the use of spoken conversation technologies for edutainment and entertainment as opposed to their use in standard information applications;
- demonstrate workable speech recognition for children's speech which is notoriously difficult to recognise with standard speech recognisers trained on adult speech-only;
- demonstrate spoken computer games, in a novel and wider sense of this term, based on a professional computer games platform; and
- create a system architecture which optimises re-use, so that it is easy to replace Andersen by, e.g., Newton, Ghandi, or the 40-some past US presidents.

The challenge of addressing domains of edutainment and entertainment rather than information systems was, in fact, chosen to make things slightly easier. Our assumption was that users of the former systems would be more tolerant to system error as long as the conversation as a whole would be

perceived as entertaining. Furthermore, the museum context-of-use requirement mentioned earlier would reduce the performance requirements on the system to those needed for 5–20 min of fun and edutaining interaction. Based on the reasoning just outlined, we chose fairytale author Hans Christian Andersen for our embodied conversational agent because of yet another pragmatic consideration. Given the need to train the system's speech recogniser with large amounts of speech data to be collected in the project, we needed a natural and convenient place to gather this data, such as the Andersen museum in his native city of Odense, Denmark, where partner NISLab is located.

1.2. Interacting with Andersen

The user meets Andersen in his study in Copenhagen (Fig. 1) and communicates with him in fully mixed-initiative conversation using spontaneous speech and 2D gesture. Thus, the user can change the topic of conversation, back-channel comments on what Andersen is saying, or point to objects in Andersen's study at any time, and receive his response when appropriate. 3D animated Andersen communicates through audiovisual speech, gesture, facial expression, body movement and action. The high-level theory of conversation underlying Andersen's conversational behaviour is derived from analyses of social conversations aimed at making new friends, emphasising common ground, expressive story-telling, rhapsodic topic shifts, balance of interlocutor "expertise" (stories to tell), etc. [2]. When Andersen is alone in his study, he goes about his work, thinking, meandering in locomotion, looking out at the streets of Copenhagen, etc. When the user points at an object in his study, he looks at the object and then looks back at the user before telling a story about the object.

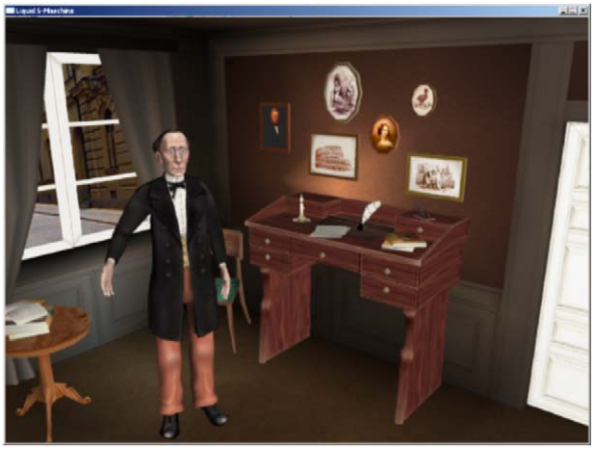

Fig. 1. Andersen gesturing in his study.

Andersen's domains of knowledge and discourse are: his works, primarily his fairytales, his life, his physical and personal presence, his study, and his interest in the user, such as to know basic facts about the user and to know which games children like to play nowadays. The user is, of course, likely to notice that Andersen does not know everything about those domains, such as whether his father actually did see Napoleon when joining his army or whether Andersen's visit to Dickens' home in England was a pleasant one. The cover story, which Andersen tells his visitors on occasion, is that he is just back and that there is still much he is trying to remember from his past.

Visiting Andersen, the user can not only talk to him, but also gesture towards objects in his study, such as pictures on the wall or his travel bag on the floor, using a touch screen. Andersen encourages his visitors to do so and has stories to tell about those objects. Using a keyboard key, the user can choose between a dozen different virtual camera angles onto Andersen and his study. The user can also control Andersen's locomotion using the arrow keys and assuming that Andersen is not presently in autonomous locomotion mode.

Some user input has emotional effects on Andersen, such as when they talk about his poor mother, the washerwoman who died early and had her bottle of aquavit to keep her company when washing other people's clothes in the Odense River. Andersen is friendly by default but he can also turn sad, as illustrated in Fig. 2, angry, such as when a child tries to offend him by asking about his false teeth, or happy, such as when the self-indulgent author gets a chance to talk about how famous he has become.

1.3. Related work

The development of the NICE Andersen system relies on several research fields, in particular those of multi-modal input systems, Embodied Conversational Agents, and interactive systems for young users.

Regarding multi-modal input, numerous prototypes have been developed for combining speech and gesture input in, e.g., task-oriented spatial applications [3], crisis management [4], bathroom

Fig. 2. Close-up of a sad Andersen.

design [5], logistic planning [6,7], tourist maps [8,9], real estate [10], graphic design [11] or intelligent rooms [12,13]. Users' multi-modal behaviour was also investigated in order to ground system development on empirical data, e.g., for the temporal parameterisation of input fusion [14].

Some general requirements to multi-modal 2D gesture/speech input systems have been proposed in standardisation efforts [15]. Unification algorithms have been applied successfully to the interpretation of task-based applications [6]. Techniques have been proposed for managing ambiguity in both the speech and the gesture modality when each of them has limited complexity, such as in [16] where different spoken commands can be combined with different gestural commands for, e.g., mutual disambiguation. Different approaches were considered for multi-modal fusion, including early fusion, which integrates signals at the feature level (for example for simultaneously training lip-reading and speech recognition), and late fusion which merges individual modalities based on temporal and semantic constraints.

One particular characteristic of the NICE Andersen system is that it offers multi-modal interaction with an animated character—a kind of interface also called Embodied Conversational Agent (ECA) [17] or Pedagogical Agent when applied to education [18]. Given the enormous challenges to achieving full human-style natural interactive communication, research on ECAs is a multi-dimensional endeavour, ranging from fine-tuning lip synchronisation details through adding computer vision to ECAs to theoretical papers on social conversation skills and multiple emotions which ECAs might come to include in the future. So far, the ECA community has put less emphasis on advanced spoken interaction than has been done in the NICE Andersen system and ECA researchers are only now beginning to face the challenges of domain-oriented conversation. Moreover, few ECA researchers have ventured into the complex territory of conversational gesture/speech input fusion.

For these reasons, we know of few ECA research systems that come close to the Andersen system prototype in being a complete demonstrator of interactive spoken computer games for edutainment and entertainment. One of the research systems closest to the Andersen system may be the US Mission Rehearsal system [19]. By contrast with the Andersen system but similar to the NICE fairytale world system, the Mission Rehearsal system is a multi-agent one, so that users can speak to several virtual agents. On the other hand, the sophisticated spoken dialogue with the Mission Rehearsal system is more task-oriented than is the conversation with Andersen; does not enable gesture and gesture/speech input; and does not target children. A few other prototypes involve bi-directional multi-modal communication and hence communication with an ECA via multi-modal input. The MAX agent [20] recognises and interprets combinations of speech and gesture, such as deictic and iconic gesture used for pointing, object manipulation, and object description in virtual reality assembly task. Combination of speech and 2D mouse gestures for interacting with a 3D ECA in a navigation task within a virtual theatre is presented in [21]. The CHIMP project had goals similar to NICE, i.e., to enable children to communicate with animated characters using speech and 2D gestures in a gaming application [22]. Similarly, some projects address fusion of users' gestures and speech when interacting with a robot. Combination of natural language and gesture to communicate commands involving directions (e.g., «turn left») and locomotion (e.g., «go over there») with a robot is described in [23]. Interaction with a humanoid robot in a kitchen scenario is described in [24]. Yet, for several of these bidirectional systems, the interaction still remains task-oriented or only addresses rather restricted conversational interaction experimentally evaluated with a children user group. The conversational dimension notably showed that turn-taking was a main issue, requiring proper output for notifying the user that the agent wants to take, keep, or give the turn.

Another domain likely to provide interesting data for the NICE Andersen project is the research on computer systems dedicated to cognitive development and child education. For example, using a simulated ECA system, Oviatt observed convergence between the spoken behaviour of children and the spoken behaviour of an animated character in a pedagogical application [25]. She also showed the differences in children's speech with this agent as compared to their speech with a human adult [26]. The effect of interacting with an agent was also observed in storytelling abilities of five-year-old girls [27]. However, neither gestural nor multi-modal children's behaviour has been studied to any great extent. Read et al. [28] studied hand-written text input from children but, to our knowledge, only [46] analysed children's multi-modal behaviour with ECAs, primarily focusing on temporal integration of speech and pen input. In this context, the evaluation of the NICE Andersen system provides more data on children's interaction with ECAs, as well as a semantic analysis of their multi-modal constructions.

1.4. Plan for the paper

In what follows, Section 2 describes the analytical steps performed prior to the design of gesture input processing as well as the specifications and algorithm of the Gesture Recogniser (GR) and the Gesture Interpreter (GI). Section 3 presents the design of the Input Fusion (IF) module. Technical and user test results on gesture-related conversation are presented in Section 4. Section 5 concludes the paper by taking a broad look at some of the challenges ahead, which have become increasingly familiar to us in the course of the work presented in this paper. Throughout, we describe the design and evaluation of the 2nd Andersen prototype, which was in part grounded on observations made on the first Andersen prototype in which the speech recognition was simulated by human wizards [29,30].

2. Gesture recognition and interpretation

2.1. Requirements on gestural and multi-modal input

In view of the richness and complexity of spoken interaction in the Andersen system, we opted for having basic and robust gesture input. Thus, gesture input has the relatively simple generic semantics and pragmatics of getting information about objects in Andersen's study, which can then be combined with the expected, richer semantics of the spoken input. We did not consider strict unification as in the task-based systems described above, as such strict semantic checking did not appear relevant in an edutainment application for children. Furthermore, the graphical on-screen objects were designed so as to avoid possible overlaps between objects in order to facilitate gesture recognition.

Fig. 3 shows the Andersen system's overall architecture, including the modules involved in gestural and multi-modal input processing: GR, GI and IF. The modules communicate via a message broker, which is publicly available from KTH [31]. The broker is a server that routes function calls, results, and error codes between modules, using TCP/IP for communication. Input processing is distributed across two input "chains" which come together in IF. Speech recognition uses

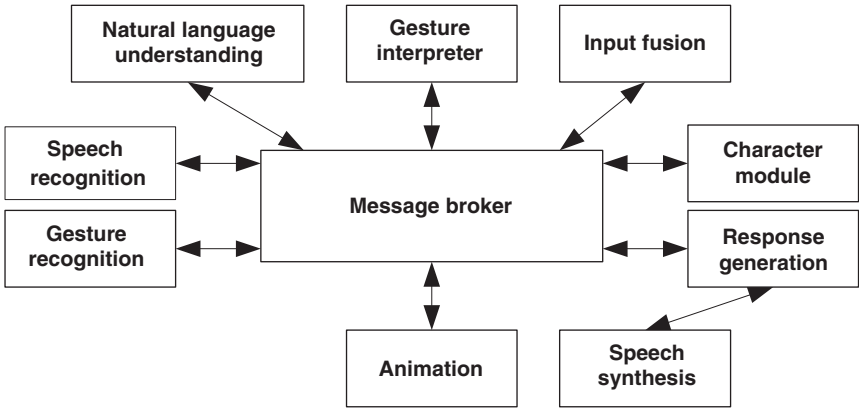

Fig. 3. General NICE Andersen system architecture.

a 1977 word vocabulary and a language model developed on the basis of three Wizard of Oz corpora and two domain-oriented training corpora collected in the project. The recogniser's acoustic models are tuned to children's voices, using approximately 70 h of data most of which has been collected in the project. A large part of this data was collected in the Odense Andersen museum, using a Wizard of Oz-simulated speech-only version of the system. The recogniser does not have barge-in (constant listening to spoken input) because of the potentially noise-filled public use environment. This restriction on the naturalness of conversation with Andersen was decided upon in order to limit the number of speech recognition errors that the system would have to deal with. Some effects are that turn-taking negotiation becomes curtailed and that the user is not able to stop Andersen from completing the story he is presently telling. It is also possible that the system would miss some backchannelling input produced by users while Andersen is speaking. Natural language understanding uses the best-recognised input string to generate a frame-based attribute/value representation of the user's spoken input, including dialogue act information. The gesture input "chain" is described in detail in the following sections.

The Andersen character module matches results produced by the IF module to potential Andersen output in context. Andersen keeps track of what he has said already and changes domain when, having the initiative, he has nothing more to tell about a domain; takes into account certain long-range implications of user input; remembers his latest output; and keeps track of repeated generic user input, including input which requires some form of system-initiated meta-communication. The character module's emotion calculator calculates a new emotional state for each conversation turn. If the input carries information which tends to change Andersen's emotional state from its default friendly state towards angry (e.g., "You are stupid"), sad (e.g., "How was your mom?"), or happy ("Who are you?"—I am the famous author Hans Christian Andersen…)—the emotion calculator updates his emotional state. If the user's input does not carry any such information, Andersen's emotional state returns stepwise towards default friendly.

Design-wise, Andersen is always in one of three output states, i.e., non-communicative action when he is alone in his study working, communicative function when he pays attention to the user's input, and communicative action when he actually responds to input. In the current system version, these three output states are not fully integrated and can only be demonstrated in isolation. The exception is when the user gestures towards an object in Andersen's study, making him turn towards the object gestured at and then turn back to face the user (the virtual camera). Response generation generates a surface language string with animation and control (e.g., camera view) tags. The string is sent to the speech synthesiser, which synthesises the verbal output and helps synchronise speech and non-verbal output, including audio-visual speech. Speech synthesis is off-the-shelf software from AT&T. Andersen's voice was chosen partly for its inherent intelligibility and naturalness, and partly for matching the voice one would expect from a 55 years old man. Finally, animation renders Andersen's study, animates Andersen, and enables the user to change camera angle and control Andersen's locomotion.

As described in the introduction, the part of the scenario related to the graphical objects displayed in Andersen's study is for the user to "indicate an object to get information about it or express an opinion about it". Table 1 lists the communicative acts identified a priori, which were likely to lead to gestural or multi-modal behaviours. The only generic gesture semantics they feature is the gestural selection of object(s) or location(s). Other possible semantics, such as drawing to add or refer to an object, or crossing an object to remove it, were not considered compatible with the NICE scenario.

A 2D gestural input has several dimensions that need to be considered by the GR/GI/IF modules: shape (e.g., pointing, circle, line) including orientation (e.g., vertical, horizontal, diagonal); points of interest (e.g., two points for a line); number of strokes; location relative to objects; input device (mouse or tactile screen); size (absolute size of bounding box, size of bounding box relative to objects); and timing between sequential gestures. Gesture processing of these dimensions is a multi-level process involving the GR, GI and IF modules. The GR computes a "low-level" semantics from geometrical features of the gesture without considering the objects in the study. The GI computes a higher-level semantics by considering the list of visible objects and their locations at the time of gesturing as sent by the object tracker from the rendering engine. Thus, the possibility that several objects are selected simultaneously cannot be

Table 1
List of identified communicative acts

	Communicative acts
1.	Ask for clarification on what to do with gesture
2.	Ask for initial information about the study
3.	Select one referenceable object
4.	Select one non referenceable object
5.	Select several referenceable objects
6.	Select an area
7.	Explicitly ask information about selected object
8.	Negatively select an object (e.g. "I do not want to have information on this one")
9.	Negatively select several objects
10.	Confirm the selection
11.	Reject the selection
12.	Correct the selection
13.	Interrupt Andersen
14.	Ask Andersen to repeat the information on the currently selected object
15.	Ask Andersen to provide more information on the currently selected object
16.	Comment on information provided by Andersen
17.	Comment on another object than the one currently selected
18.	Select another object while referring to the previous one
18.	Select another object of the same type than the one currently selected
20.	Move an object (user may try to do that although not possible and not explicitly related to the user's communicative intention)
21.	Compare objects
22.	Thank

detected by the GR and has to be detected by the GI. The IF computes a final interpretation of gesture by combining the GI output with the Natural Language Understanding (NLU) output.

In the test of the 1st Andersen prototype, some users made several sequential gestures (e.g., parts of a circle) on the same object, which might be due to the fact that the gesture stroke was not highlighted on the screen (which might be due to insufficient finger pressure on the touch screen or a faulty touch screen setting), that Andersen would not give any feedback, such as gazing at the gestured object, or that their finger simply slipped on the tactile screen. This resulted in duplicated messages sent by the GI and thus to output repetitions by the system. In order to avoid this, we decided to have the GI group several sequential strokes on the same object as a single gesture on this object.

Other difficulties include the facts that some objects have overlapping bounding boxes some of which may be partly hollow, such as for the coat-rack, and that some objects are partly hidden by other objects as when, e.g., a chair is behind the desk from several viewpoints.

Table 2
Definition of GR output classes

GR output class	Features of input gesture (shape and size)
Pointer	Point. Very small gesture (10×10 pixels) of any shape including garbage Very small line, tick, scribble
Surrounder	The following "Surrounding" gesture shapes (for single object selection) were logged during Prototype-1 user tests and are used for training the GR: • Circle, open circle, noisy circle, vertically/ horizontally elongated circle • "alpha", "L", "C", "U"-like gestures with symmetrical shapes • Square, diamond, vertical/ horizontal rectangle
Connect	Vertical, Horizontal, Diagonal lines. Multiple back-and-forth lines
Unknown	Garbage gesture. The bounding box is not very small (otherwise recognised as a point)

2.2. Gesture recognition

The gestural analysis described above resulted in the set of shapes described in Table 2.

As a result of gesture recognition, the GR sends to the GI a «grFrame» including the 1st best gesture shape recognised. The two-stroke "cross" shape is recognised when two crossing lines are drawn. It is recognised by the GI (instead of the GR) in order to avoid confusing the delay between the two strokes of the cross with the delays between different gestures. If the multi-stroke gestures were to be recognised by the GR, the GR would have to delay the sending of recognised lines to the GI as, e.g., the GR would wait for the second line of the cross. This delay would add to the delay in the GI for grouping sequential gestures of any type on the same object. In order to avoid this sum of delays, we decided to have multi-stroke gestures recognised by the GI since, there, the delay is used both for waiting for (1) a possible 2nd stroke of a multi-stroke gesture and (2) another single-stroke gesture on the same object.

When a gesture is detected by the GR, a «startOfGesture» message is sent by the GR to the

IF before launching shape recognition in order to enable appropriate timing behaviour in the IF. When the GR is not able to recognise the shape or when the user makes noisy gestures, the GI can try to recover, considering them as surrounder gestures, and hopefully detect any associated object. The goal is to reduce the non-detection of gestured objects. Indeed, surrounder gestures logged during Prototype-1 evaluation were quite noisy and included contours of objects. Another possibility would have been to induce the user to gesture properly and not to forward unknown shapes to the GI, but that was considered inappropriate for a conversational application for children. The GR also sends the gesture bounding box to the GI.

The GR uses a back-propagation neural network trained with gestural data logged from Prototype-1. Training involves several steps: manual labelling of logged shapes, training of the neural network, and testing and tuning its parameters. The general algorithm of the GR is shown below.

```
Algorithm GR
  When a gesture is detected:
      Send a ''startOfGesture'' message to IF

      If the bounding box of the gesture is very small (10 × 10)
      Then set shape = ''pointer''
      Else
            Convert the gesture points to a slope feature array.

            Test the feature array with the neural network.

            set shape = result from the neural network
            (either ''surrounder'' | ''connect'' | ''unknown'')

            If the shape is ''connect''
            Then compute start and end points of the line

      Build a grFrame for this newly detected gesture

      Send the grFrame to the GI

End of Algorithm GR
```

2.3. Gesture interpretation

The GI module aims at detecting the object(s) the user gestures at. It has been designed by considering the properties of the graphical objects that are displayed and which the user is able to refer to. The properties are:

- spatial ambiguities due to objects that have overlapping bounding boxes, or objects that are in front of larger objects, such as the objects on Andersen's desk;
- the singular/plural affordance of objects, e.g., a picture showing a group of people might elicit either singular spoken deictics, such as «this picture», or plural spoken deictics («these people»);
- perceptual groups which might elicit multiple-object selection with a single gesture, or for which a gesture on a single object might have to be interpreted as a selection of the whole group, such as the group of pictures on the wall [32].

Following gesture interpretation, the GI sends a «giFrame» to the IF module. This frame includes one of the three attributes "select" (a gesture on a single object), "reference ambiguity" (several objects were gestured at), or "no object" (a gesture was done, but no associated referenceable object could be detected), as defined in Table 3. Gesture recognition confidence scores are not considered since a fast answer from the character is preferred over an in-depth resolution of ambiguity in order to enable fluent conversation. Moreover, due to the challenging complexity in recognising children's conversational speech, it was preferred to ensure robust gesture interpretation by avoiding, as far as possible, overlaps between graphical objects. Such design choices wrt. to the graphical environment enabled us to reach high-accuracy recognition of gestured objects during monomodal tests held prior to the test involving multi-modal fusion and children users. Indeed, as will be described in the

Table 3
Definition of GI output classes

GI output semantic class	GR output class	Graphical context
Select	Pointer Cross Surrounder Connect	Gesture bounding box overlaps with bounding box of only one object.
	Sequential: Pointer Cross Surrounder Connect	On the *same* object (close in time).
referenceAmbiguity	Surrounder Cross Connect Sequence of pointers or other shapes than unknown	Bounding box of gesture overlaps with the bounding boxes of several objects.
noObject	Any except unknown	GI failed to detect any object although a gesture was made by the user (gesture on empty space; selection of non referenceable objects).

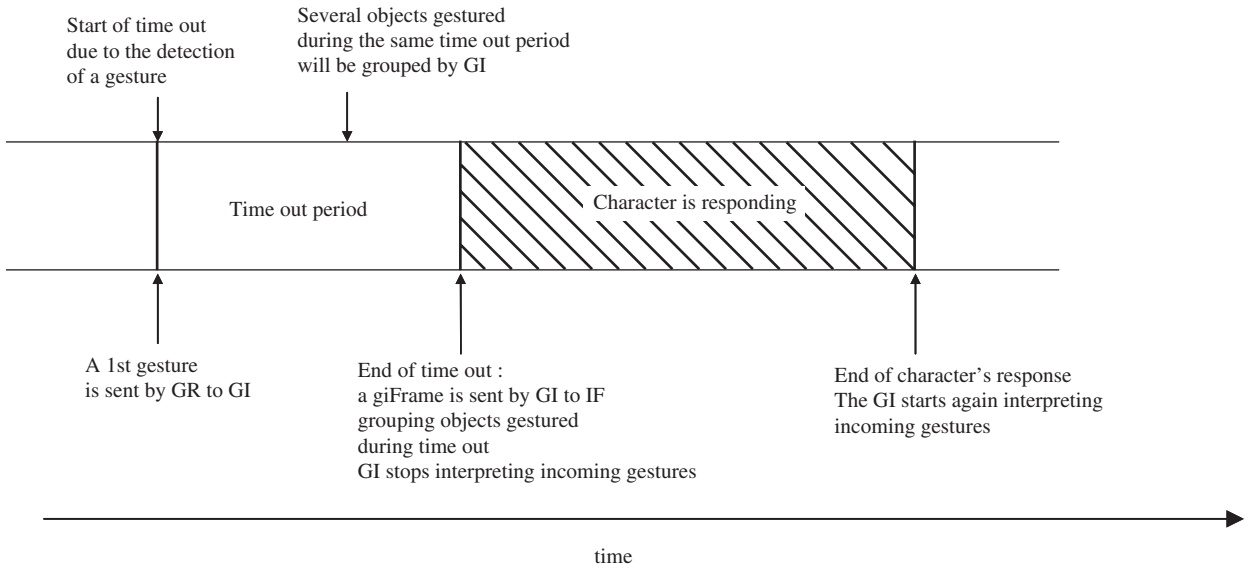

Fig. 4. Temporal management in the GI module.

section on evaluation, assigning scores to results of gesture interpretation would not have addressed the problems observed in the management of multimodal behaviour.

The conversational context of the Andersen system requires management of timing issues at several levels (Fig. 4). In order to avoid endless buffering of the user's input while Andersen is speaking, gesture interpretation is inhibited during preparation and execution of Andersen's verbal and non-verbal behaviour. In order to sequentially group objects gestured at, the GI has a relatively fast timeout. It collects what it gets before the timeout and then passes it on to the IF. The message sent by the GI to the IF may include reference to one or several objects. If several objects are referenced, this may mean either that a single gesture was done on several objects or that

sequential gestures were done on different objects. An object does not appear twice in the giFrame even in the case of multiple gestures on the same object. The GI collects references to one or several objects in the given time window and passes them to the IF as a single gesture turn. The timeout period is reset each time a new gesture is recognised.

The 2nd Andersen prototype requires that once the timeout has been started, incoming gestures are ignored by the GI. The Character Module notifies the GI with an «EndOfBehavior» message that Andersen has finished his verbal and nonverbal output turn, so that the GI can start interpreting gestures again. The same notification is sent to the speech recogniser. The GI timeout is analogous to the lack of barge-in in the speech recogniser. However, the GI timeout may be less of a restriction on the naturalness of conversation since few users tend to do 2D touch screen gesture without speaking.

The following durations were selected as default values for the GI module:

- timeout period duration: 1.5 s. This is compatible with observations made during the Prototype-1 user tests;
- maximum duration of waiting for the character's response = 6 s. After this the GI starts interpreting gestures again.

These specifications resulted in the design of the following algorithm for time management in the GI:

```
Algorithm GI
Input: incoming messages from GR and CM
Output: messages sent by GI to IF
Variable: list of object names gestured during timeout

{Processing of an incoming grFrame from GR}
If a grFrame is received from GR
Then
      If the character's response is currently pending
      Then
            Ignore grFrame
      Else
            If gesture time out period is not started
            Then start gesture time out period

            Call bounding box algorithm to detect objects

            Store name of detected object(s)
            in the list of gestured objects (avoid duplicates)

{Gesture time out period has finished}
If end of timeout period
Then
      If no object was detected during timeout
      Then
            Build a ''noObject'' giFrame


      If a single object has been detected during timeout
      Then
            Build a ''select'' GIFrame with name of this object

      If several objects have been detected during timeout
      Then
            Group objects names in a ''referenceAmbiguity'' GIFrame


      Send the GIFrame to IF
      Set characterResponsePending to true
```

```
{Character's response is finished}
If      message is ''EndOfBehavior'' is received from the Character/Dialog Module OR
        message ''EndOfBehavior'' has been waited for too long
Then
        Set characterResponsePending to false
        Set gesture detection period not started
        Enable GI to start new timeout if a gesture is detected

End of Algorithm GI
```

In 3D graphics, some objects hide others, such as when a vase is hiding a table. Yet, the graphical application only delivers the coordinates of all the objects, which are partly in the camera viewpoint without informing the GI if these objects are hidden or not by some other visible objects. The objects which are hidden must not be selectable by gesture, even if the gesture is spatially relevant. In the bounding box algorithm, we used the depth (Z dimension) of the closest side of the bounding box of objects to compute hidden objects. The salience value computed for each object is weighted by a factor of the distance, which is maximal when the front of the object is near the camera and decreases quickly for objects which are far from the camera. Yet, an object closer on its Z-dimension can actually be partially hidden by one further away, such as a vase on a table, which hides the part of the table, which is behind the vase. Thus, the size of the object is also considered in the algorithm. An object, which better fits the size of the gesture is more likely to be selected.

3. Input fusion

3.1. Requirements and specifications of input fusion

IF in the Andersen project aims at integrating children's speech and 2D gestures when conversing with virtual characters about 3D objects. In principle, IF is subject to some general requirements to multimodal input systems, such as the need to manage and represent timestamps of input events, multi-level interpretation, composite input, and confidence scores [15]. Yet, the conversational goal of the system and the fact that it aims at being used by children make it different from current research on systems which use speech and gesture for task-oriented applications as described in the introduction.

Both speech-only input and gesture-only input can be semantically and pragmatically independent. In other words, using either, the user can input a complete communicative intention to the system. As for combined gesture and speech in an input turn, their relationship regarding the semantics of object selection may be of several different kinds. Thus, the input speech may be either (i) redundant relative to the input gesture as in ⟨pointing at the picture of Andersen's mother⟩ "Tell me about your mother", (ii) complementary to the input gesture as in ⟨pointing at object⟩ "What is this?", (iii) conflicting with the input gesture as in ⟨pointing at the picture of Andersen's mother⟩ "Tell me about your wife", or (iv) independent of the input gesture as in ⟨pointing at the feather pen⟩ "Do you live here?".

Given the formal patterns of relationship between speech and gesture input just described, it would appear that speech-gesture IF is required in the two cases of redundancy and complementarity. Conversely, IF is excluded in all cases of speech-gesture independence, i.e., speech-only input, gesture-only input, and independent, but simultaneous speech and gesture inputs. When independent gesture and speech occur at the same time, the system should not merge them. As for speech/gesture conflict, we decided to trust the gesture modality, as it is more robust than the speech recognition in the context.

The IF module integrates the messages sent by the NLU and the GI modules and sends the result to the character module. The IF parses the message sent by the NLU to find any explicit object reference (e.g., "this picture") or implicit reference (e.g., "Jenny Lind?", "Do you like travelling?") which might be integrated with gestures on objects in the study. In order to do so, the IF parses the frame produced by the NLU and spots the following concepts: object in study, fairytale,

Table 4
Description of multimodal sequences observed in the Prototype-1 video corpus

Succession of modalities	Delay[a] between modalities (s)	Object gestured at	Shape of gesture	Spoken utterance + NLU frame	Cooperation between modalities
Gesture–speech	2	Picture of Coliseum	Circle	"What's this?"	Complementarity
Simultaneous	0	Picture of Andersen's mother	Circle	"What's that picture?"	Complementarity
Simultaneous	0	Hat	Circle	"I want to know something about your hat."	Redundancy
Gesture–speech	4	Statue of 2 people	Circle	"Do you have anything to tell me about these two?"	Complementarity
Simultaneous	0	Statue of 2 people	Point	"What are those statues?"	Complementarity
Gesture–speech	4	Picture above book-case	Circle	"Who is the family on the picture?"	Complementarity
Gesture–speech	3	Picture above book-case	Circle	"Who is in that picture?"	Complementarity
Simultaneous	0	Vase	Circle	"How old are you?"	Concurrency

[a]The delay between modalities was measured between end of first modality and end of second modality.

fairytale character, family, work, friends, country, and location. It produces messages containing the "fusion status" which can be either "ok", i.e., the utterance and the gestured object were integrated because a reference was detected in the NLU message and in the GI; "none", i.e., the utterance and the gesture were not integrated either because there was only one of them, or because the IF could not decide if they were consistent or not regarding the number of references to objects in speech and gesture; or "inconsistent", i.e., the utterance and the gesture were inconsistent regarding the number of referenced objects. In case of successful integration, the semantic representation of gesture (the detected object(s)) is inserted into the semantic representation sent by the NLU. The IF module also manages temporal delays between gesture and speech via several timeouts and messages signalling start of speech and start of gesture.

The IF specifications described above were driven by a conversation analysis that generated a set of 233 multi-modal combinations which users might produce. This set includes the multi-modal behaviours observed during the Prototype-1 user tests.

3.2. *Multi-modal behaviours in the Prototype-1 user tests*

During the Prototype-1 user tests, 2 h were videotaped (about 22% of the tests). Only 8 multi-modal behaviours were observed. These are shown in Table 4.

These examples provide illustrative semantic combinations of modalities:

- Deictic: "What's this?" + circling gesture on the picture of the Coliseum.
- Type of object mentioned in speech: "What's that picture?" + circling gesture on the picture of Andersen's mother; "I want to know something about your hat" + circling gesture on the hat.
- Linguistic reference to concepts related to the graphical object (e.g., "dad" and gesture on a picture) instead of direct reference to the object type or name ("picture");
- Incompatibility between internal singular representation of objects and their plural/singular perceptual "affordance", e.g., a single object is referred to in the user's speech as a plurality of objects: "Do you have anything to tell me about these two?" (or "What are those statues?") with a circling gesture on the statue of two characters which are internally represented as a single object.

Several objects might elicit such plural/singular incompatibility. They visually represent several entities of the same kind, but they are (system-) internally represented as a single object. They could be thus referred to as a single object or as several objects, their number being foreseeable for some of them: books (number >2); boots (2); papers (>2); pens (2); statue (2).

Conversely, although this was not observed as such in the Prototype-1 user test video, several objects of similar type and in the same area might be

perceived as a single ''perceptual group'' [32] and might elicit a plural spoken reference combined with a singular gesture on only one of the items in the group: the group of pictures on the wall above the desk, the ''clothes group'' (coat–boots–hat–umbrella), the furniture (table and chairs), the small objects on the small shelf.

3.3. Temporal dimension of input fusion

A main issue for IF is to have a newly detected gesture wait for a possibly related spoken utterance. How long should the gesture wait before the IF decides that it was indeed a mono-modal behaviour? We decided to use default values for delays to drive the IF to have gestures wait a little for speech (3 s) and have speech wait for gesture for a very short while only, since this is compatible with the literature [33] and the Prototype-1 user tests observations. We have also introduced management of ''StartOfSpeech'' and ''StartOfGesture'' messages sent to the IF in order to enable adequate waiting behaviour by the IF. Four temporal parameters of the IF have been defined to answer the following questions:

- How long should an NLU frame wait in the IF for a gesture when no ''StartOfGesture'' has been detected (Speech-waiting-for-gesture-short-delay)? The default value is 1 s.
- How long should an NLU frame wait in the IF for a gesture when a ''StartOfGesture'' has been detected (Speech-waiting-for-gesture-long-delay)? The default value is 6 s.
- How long should a GI frame wait in the IF for a NLU frame when no StartOfSpeech has been detected (Gesture-waiting-for-speech-short-delay)? The default value is 3 s.
- How long should a GI frame wait in the IF for an NLU frame when StartOfSpeech has been detected (Gesture-waiting-for-speech-long-delay)? The default value is 6 s.

The part of the IF algorithm that manages temporal behaviour is specified with the instructions to be executed for each event that can be detected by the IF: a new NLU frame is received by the IF, a new GI frame is received by the IF, a ''StartOfSpeech'' message is received by the IF, a ''StartOfGesture'' message is received by the IF, a ''Speech-waiting-for-gesture'' times out, and a ''Gesture-waiting-for-speech'' times out.

The IF behaviour is described informally below for each of these events.

Init()

{Starts with ''short'' delays when no start of speech or gesture has been received. When start of speech/gesture will be received, these will be set to longer delays since there is a very high probability that an associated speech or gesture frame will be received afterwards by the IF}

```
Speech-waiting-for-gesture-delay = Speech-waiting-for-gesture-short-delay
Gesture-waiting-for-speech-delay = Gesture-waiting-for-speech-short-delay
```

When a new NLU frame is received by the IF

```
{Test if a gesture was already waiting for this NLU frame}
If the timeout Gesture-waiting-for-speech is running
Then
      {A GI frame was already waiting for this NLU frame}
      Call semantic fusion on the NLU and the GI frames
      Stop-Timer(Gesture-waiting-for-speech)
Else
      {This new NLU frame will wait for incoming gesture}
      Start-Timer(Speech-waiting-for-gesture)
```

When a new GI frame is received by the IF

```
{Test if a NLU frame was already waiting for this GI frame}
If the timeout Speech-waiting-for-gestureis running
Then
```

```
        {A NLU frame was already waiting for this GI frame}
        Call semantic fusion on the NLU and the GI frames
        Stop-Timer (Speech-waiting-for-gesture)
Else
        {This new GI frame will wait for incoming speech}
        Start-Timer (Gesture-waiting-for-speech)
```

When a *startOfSpeech* message is received

```
{A new NLU frame will soon arrive. Ensure that the GI frame that is already waiting
waits longer or that if a new GI frame arrives soon (since a StartOfGesture was
received) it will wait for the NLU frame}

Gesture-waiting-for-speech-delay = Gesture-waiting-for-speech-long-delay

If Gesture-waiting-for-speech is running
Then
      Restart-Timer (Gesture-waiting-for-speech)
```

When a *startOfGesture* message is received

```
{A new GI frame will soon arrive. Ensure that the NLU frame that is already waiting
waits longer or that if a new NLU frame arrives soon (since a StartOfSpeech was
received) it will wait for the GI frame}

Speech-waiting-for-gesture-delay = Speech-waiting-for-gesture-long-delay

If Speech-waiting-for-gesture is running
Then
       Restart-Timer(Speech-waiting-for-gesture)
```

When timeout *Speech-waiting-for-gesture* is over

```
{A NLU frame has waited for a GI frame which did not arrive}
Build and send an IF frame containing only the NLU frame
Stop-Timer(Speech-waiting-for-gesture)
Init()
```

When timeout *Gesture-waiting-for-speech* is over

```
{A GI frame has waited for a NLU frame which did not arrive.}
Build and send an IF frame containing only the GI frame
Stop-Timer(Gesture-waiting-for-speech)
Init()
```

3.4. Semantic dimension of input fusion

Regarding semantic IF we have decided to focus on (1) the semantic compatibility between gestured and spoken objects, and (2) the plural/singular property of these objects. We limited ourselves to one reference per NLU frame and identified 16 possible semantic combinations of speech and gesture (Table 5).

Only cases 11, 12, 15, and 16 can possibly lead to fusion in the IF, as described above. We systematically analysed each of the 16 cases. Below, we specify the instructions to be executed by the IF and the output it produces for each case. The instructions consider the following features of speech and gesture references: singular/plural, reference/no reference, semantic compatibility.

Semantic compatibility between gestured and spoken objects is evaluated by the IF via semantic distance computation which is less strict than object type unification and was expected to be more appropriate for conversational systems for children. Semantic distance computation makes use of a graph of concepts connected with an "is-related-to" relation. Each concept is represented by: a name

Table 5
Analysing 16 combinations of speech and gesture along the singular/plural dimension of references (only cases 11, 12, 15, and 16 can possibly lead to fusion in the IF)

GI/NLU "referenceAmbiguity"	No message from GI	1 message from GI "noObject"	1 object detected by GI "select"	Several objects detected by GI
No message from NLU	1	2	3	4
1 message from NLU but no explicit reference in NLU frame	5	6	7	8
1 message from NLU with 1 singular reference	9	10	11	12
1 message from NLU with 1 plural reference	13	14	15	16

(e.g., "feather Pen", "_Family"), a plural Boolean (e.g., "true" for the statue of two people), a singular Boolean (e.g., "true" for the feather Pen), a Boolean describing if it is an object in the study ("pictureColiseumRome") or an abstract concept ("_Mother"), and the set of semantically related concepts (generic relation "isRelatedTo").

A reference detected by the NLU module is represented in the IF by: a Boolean stating if it is solved, a Boolean stating if it is plural/singular, and a Boolean stating if it is numbered (if yes, an attribute gives the number of referenced objects, e.g., "two" in the reference "these two pictures").

A perceptual group is represented by the same attributes as a single concept, and by the set of concepts, which might be perceived as a group (e.g., the set of pictures above the desk).

The identified cases of semantic combination described above are integrated in a single algorithm for semantic fusion. The informal algorithm below only details cases for which one message has been sent by the NLU and one by the GI, i.e., cases 6–7–8, 10–11–12, 14–15–16 in our analysis.

After IF, when required, an IF frame is sent to the character module. An attribute called "fusion Status" is used in the IF frame to indicate if the input was monomodal ("none"), successful ("ok") or unsuccessful ("inconsistency"). Gestures towards objects that cannot be referenced are ignored and hence are not passed to the character module.

Algorithm Semantic Fusion (NLU frame, GI frame)
{Manage each multimodal combination case. We suppose that one NLU frame and one GI frame have been received by the IF}

```
IF there is no explicit reference in the NLU frame
THEN  {CASES 6-7-8}
      Group both frames
      Send them to the Character Module with a fusion status set to none
ELSE
      IF there is only one reference in the NLU frame
      THEN
            IF the reference is singular
            THEN call Semantic Fusion Singular NLU (NLU frame, GI frame)
            ELSE call Semantic Fusion Plural NLU (NLU frame, GI frame)
```

Semantic Fusion Singular NLU (NLU frame, GI frame)
{The referential Expressionin the NLU frame is singular:
CASES 10-11-12 (not perceptual group)}

```
IF there is at least one object selected by GI,
which is semantically compatible with the NLU reference
THEN
```

```
      {Do semantic fusion (possibly not considering plural constraint
      if there was several gestured objects)}
      Resolve the NLU reference with the compatible gestured object(s)
      Send the modified NLU frame to the Character Module
ELSE
      {No gestured object revealed compatible with the NLU reference}
      Signal inconsistency
      Send NLU frame and GI frame to the Character Module
```

Semantic Fusion Plural NLU (NLU frame, GI frame)

```
{The Referential Expression is plural: CASES 14-15—16—12 (perceptual group)}
IF more than one object from GI is semantically compatible with the NLU reference
THEN
      {Do semantic fusion}
      Resolve the plural NLU reference with the compatible gestured object(s)
      Send the modified NLU frame to the Character Module
ELSE
      {Manage perceptual groups}
      IF there is only one object from GI compatible with NLU reference
      and this object belongs to a perceptual group
      THEN
      {Do semantic fusion}
          Resolve the plural NLU reference with the perceptual group of objects
          Send the modified NLU frame to the Character Module

      ELSE
          IF   the GI object is compatible with the NLU reference
               but does not belong to a perceptual group

          THEN  {Do semantic fusion (not considering plural constraint)}
                 Resolve NLU reference with the compatible gestured object
                 Send the modified NLU frame to the Character Module

          ELSE   {No gestured object compatible with the NLU plural ref.}
                 Signal inconsistency ; Send NLU frame and GI frame
```

The different feedforward and feedback mechanisms that have been implemented to enable proper coordination of multi-modal input with Andersen's behaviour are summarised in Fig. 5.

3.5. Character module processing

Given the many design-time uncertainties concerning how children would use combined speech and gesture input, we chose a simple processing scheme for gesture-related input in the character module. The IF frame goes to the character module's conversation mover, which tries to match the input to candidate system output. The conversation mover passes on its results to the conversation mover post-processor whose task it is to select among the conversation mover outputs a single output candidate to pass on to the move processor which analyses the candidate in the discourse history and domain knowledge contexts. The conversation mover does nothing about gesture-related input, i.e., gesture-only input and combined gesture-speech input, but simply passes them on to the conversation mover post-processor.

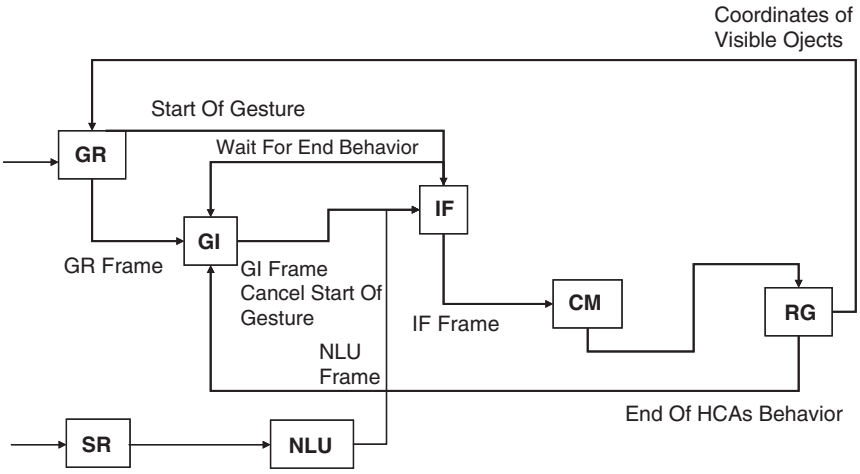

Fig. 5. Feedforward and feedback messages for managing multi-modal input conversation with Andersen (GR = Gesture Recogniser, GI = Gesture Interpreter, SR = Speech Recogniser, NLU = Natural Language Understanding, IF = Input Fusion, CM = Character Module, RG = Graphic Renderer). Messages "GRFrame", "GIFrame", "NLUFrame", and "IF Frame" enable the transmission of processing results of modules. Messages "StartOfGesture", and "CancelStartOfGesture" enable proper management of temporal relations between speech and gestures. Messages "WaitForEndBehavior" and EndOfHCABehavior" enable inhibition of gesture processing while the character is responding, hence regulating turn-taking. Message "GIFrame" is also used by the character to provide gaze feedback on the gestured object.

Informally, the post-processor's algorithm for gesture-related input is:

- check if multiple labels include label(s) prefixed by g_ [these are gesture object labels]
  - if yes, remove all labels **not** prefixed by g_
    - if only one label remains, send label to move processor END
    - if several labels remain, continue
- randomly select a label among the multiple labels left and send the selected label to move processor END

Thus, the character module ignores the "inconsistency" label from the IF and does not attempt to produce meta-communication output in an attempt to resolve the inconsistency claimed by the IF. We selected this solution because of the problems we have identified with singular vs. plural deictic expressions and what they might refer to (cf. Section 4). Furthermore, the character module does not process the spoken input in cases where the IF has deemed IF to be "ok". Also, by not processing the spoken input in cases of independent concurrency, i.e., when the user points to some object(s), but speaks about something else entirely, the strategy adopted means that Andersen at least manages to address one of the user's concerns, i.e., that of getting a story about a referenceable object. What he does not do is keep in mind that the user had spoken about something else entirely whilst pointing to some object(s). Our design reasoning was that the user, when noticing this, might simply come back and repeat the spoken input in a subsequent turn. Arguably, this design decision is an acceptable one since the user (i) does get a reply wrt. to the object pointed to and (ii) has ample opportunity to come back to the unrelated issue posed in the spoken part of the input. Given the overall design of the Prototype-2 system, the only apparent flaw would seem to be the fact that the user's spoken input might relate more closely to gesture input information randomly *discarded* by the post-processor than to the gesture input information randomly *chosen* by the post-processor. However, selecting wisely in this situation would either (i) require the conversation mover to have contextual knowledge which it does not possess or (ii) that the post-processor forward multiple output candidates to the move processor which does have contextual knowledge, and this is not possible in the Andersen Prototype-2 system.

4. Evaluation

4.1. Methodology

The Prototype-2 Andersen system was tested with 13 users (six boys and seven girls) from the target

user population of 10–18 years old children and teenagers in February 2005. All users were Danish school kids aged between 11 and 16 and with an average age of 13 years. Their English skills were not rated prior to the test as we wanted to test the system with a random sample of target users. The Prototype-1 test—following which the 18 children users' English skills were rated for speech recogniser training purposes—had shown that Danish kids are generally able to conduct conversation with Andersen even though, of course, their English proficiency varies significantly depending upon factors, such as age, individual differences, and temerity in addressing Andersen in the presence of unfamiliar adults. As in the Prototype-1 user test, in the test of Prototype-2 only a single child had significant difficulties carrying out conversation with Andersen. In the post-test structured interview, the users were asked about their knowledge of Andersen's fairytales. Their responses were all rated by two independent raters at 2 on a 3-point scale, which corresponds closely to the findings in the post-test interview following the Prototype-1 test. Danish children generally have substantial knowledge about Andersen's fairytales. Only two of the Prototype-2 users had had conversation with Andersen before, i.e., in the Prototype-1 user test.

The test was a controlled laboratory test rather than a field test in the Andersen museum. For the first user test of a strongly modified second prototype, it is often preferable to make use of the laboratory environment in order to be able to fully control the conditions of interaction, such as advance notice of users in order for them to plan for the entire (60–75 min) duration of the test which included structured post-trial interviews, common instructions to all users for each test phase, timing of the two different test conditions that were used for all users, etc. Admittedly, a field trial would have provided more realistic data on system use, but this data would also have been very different from the data collected in the lab.

Users were wearing a microphone/loudspeaker headset. They used a touch screen for gesture input and a keyboard for controlling virtual camera angles and for controlling Andersen's locomotion. Each user had a total of 35 min of multi-modal interaction with Andersen, the conversation being conducted in English. Each user interacted with the system in two different test conditions. In the first condition, they received basic instructions on how to operate the system but not on how to speak to it, and then spent approx. 15 min exploring the system through conversation with Andersen. In the second condition, in order to steer the users through a cross-section of Andersen's domain knowledge and put pressure on the system's ability to handle substantial user initiative in conversation, they received a handout with 11 issues they might wish to address during conversation at their leisure for 20 min, such as "Try to offend Andersen" or "Tell Andersen about the games you like to play". Fig. 6 shows a user in action.

Fig. 6. A user talking to the 2nd Andersen system prototype.

Two cameras captured the user's behaviour during interaction and all main module outputs were logged. Following the test, each user was interviewed separately about his/her experience from interacting with Andersen, views on system usability, proposals for system improvements, etc.

4.2. Comparative analysis of video and log files

Eight hours of interaction were logged and captured on video. In order to evaluate the GR, GI and IF modules, the gesture-only and gesture-combined-with-speech behaviours were analysed based on the videos and the log files. The videos were used to annotate the real behaviours displayed by users in terms of: spoken utterances related to gestural behaviour, the objects gestured at (including each non-referenceable object, i.e., objects in Andersen's study for which the animation does not have an id to forward to the GI), and obvious or possible misuse of the tactile screen in case the corresponding gesture was not detected by the GR. The log files were used to check the output of each

module, to compare the output to the observed behaviour from the video, and to classify reasons for, and cases of, failure.

We made a distinction between the success of the interaction and the success of the processing done by the gesture and multi-modal modules. *Multi-modal interaction* was considered successful if the system responded adequately to the user's behaviour, i.e., if the character provided information about the object the user gestured at and/or spoke about. *Module success* was evaluated by comparing the user's behaviour and the output produced by the modules in the log files. In some cases, the interaction was successful although the output of the module was incorrect, implying that the module error was counter-balanced by other means or modules. In some other cases, the interaction was unsuccessful although the output of the module was correct, implying that an error occurred in some other module(s). Interaction success for multi-modal input provides information on, among other things, the use of inhibition and timing strategies which enable proper management of some redundant multi-modal cases via the processing of only one of the modalities.

4.2.1. Gesture recognition

281 gesture shapes onto the tactile screen were logged. The shapes were manually labelled without displaying the result of GR processing (blind labelling). To enable fine-grained analysis of gesture shapes, the labelling made use of 25 categories of shapes. We found that 87.2% (245) of the logged gestures had been assigned the same category by the GR and by the manual labelling process. The fine-grained categories reveal a high number of diagonal lines ($90/281 = 32\%$) and explicitly noisy categories ($44/281 = 16\%$), such as garbage, noisy circle, and open circle of various orientations. The distribution of shapes in the GR and the manual labelling are similar.

4.2.2. Gesture interpretation

As observed in the videos, the users made 186 gesture-only turns. If we use the number of IF frames (957) for counting the number of user turns—this is not exact as sometimes a single spoken turn might be divided into several recognised utterances—gesture-only turns correspond to 19% of the user turns.

One hundred and eighty-seven messages were produced by the GI module. By comparing the log files and the videos, we found that 54% of the user gestures led to a GI frame, 30% were cancelled because detected after GI timeout and during or before the character's response, and 16% were grouped because they were done on the same object.

The repartition of the gesture interpretation categories is the following: $125/187 = 67\%$ detected a single referenceable gestured object, $61/187 = 33\%$ did not detect any referenceable object, and only one detected several referenceable objects in a single gesture. One multi-object gesture was observed in the video, but this gesture included one referenceable object and two non-referenceable objects and was thus interpreted as selection of a single object by the system.

Fifty one percent of the gesture-only behaviours led to interaction success. The reasons for the 49% cases of interaction failure were classified as follows: gesture on non-referenceable objects (62%), gesture during GI inhibition (17%), system crash (14%), unexplained (4%), gestured object not detected (2%), gesture not detected (1%). Most of the interaction failures (76%) were thus due either to gestures onto non-referenceable objects or to input inhibition. On average, each user gestured at 11 referenceable objects and 4 non-referenceable objects.

4.2.3. Input fusion

As observed in the videos, the users made 67 multi-modal turns combining gesture and spoken input. If we use the number of IF frames as our number of user turns, multi-modal turns correspond to 7% of the user turns. Among the 957 messages logged by the IF, only 21 (2%) were processed by the system as multi-modal constructions.

Seventy percent of the multi-modal turns were produced in the first test condition, cf. Section 4.1. This is the same proportion as for gesture-only behaviours. It is probable that, during the first test phase, the users explored the 3D environment, testing objects by gesturing and sometimes speaking at the same time to find out if Andersen had stories to tell about those objects. When the second test condition started, the users had already received information about a number of objects and preferred to address topics other than the objects in the study. In support of this interpretation it may be added that only one of the 11 issues in the second-condition hand-out concerned objects in Andersen's study (cf. Section 4.1).

Regarding the users' multi-modal behaviours, we also analysed interaction success and IF success. In 24 multi-modal turns, the IF was unsuccessful, but interaction was successful. Sixty percent of the multi-modal behaviours led to interaction success. Analysis of the output of the IF module reveals that it worked well for 25% of the multi-modal cases. It is quite difficult to compare such results with the literature since there are very little experimental results on multi-modal fusion in conversational applications for children. For example, Kaiser et al. [16] observed an overall success in functional accuracy of 59.1% and 81.4% for multi-modal recognition but during adult's speech and 3D gestures multi-modal commands for manipulating 3D objects.

The reasons for failure of processing multi-modal behaviours were collected from the video and log files and are listed in Table 6.

A closer analysis was done of the many "timer too small" cases, i.e., the cases in which the IF's 1.5 s waiting time for linguistic input after having received gesture input from the GI, was not long enough. The linguistic input did arrive and was temporally related to the gesture input, but it arrived too late for IF to take place, the gesture input already having been sent to the character module. In 85% of these 21 cases, the timestamp of the IF's "StartOfSpeech" message was evaluated as being incorrect compared to the start of speech observed in the video. It would have been inappropriate to have the user wait for such a long period, e.g., 10 s in several cases. For example, the "start of speech" would be logged as arriving in the IF 14 s after the "start of gesture" although, in the video, the user starts to speak only 1 s after the start of gesture. Indeed, given the limited semantics of gesture involved, i.e., only selection of objects, and the frequent redundancy of speech and gesture in the conversational context, the strategy to take an early decision for gesture-only behaviour enabled us to obtain 60% of interaction success for multi-modal behaviour while avoiding the user waiting too long for the system's response. The IF would briefly wait for NLU input and then send its frame to the character module, ignoring any delayed NLU input. The explanation for the delayed "start of speech", as this is labelled by the IF, turned out to be a flaw in the speech recogniser's detection of *end* of speech, so that the recogniser would continue to listen until timeout even if the user had stopped speaking maybe 10 s before. This flaw turned out to be more complex to correct than expected because it was due to the fact, unknown to us at the time, that we should have used a different approach for implementing end of speech detection in the Scansoft recogniser.

Table 6
Reasons of failure in processing of multimodal behaviours

	NB	%
Timer too small	21	43
Speech recognition error	9	18
Input inhibited	6	12
Not a referenceable object	4	8
Gesture not detected	4	8
System crash	2	4
Unexplained	2	4
Gestured object not detected	1	2
Total	49	100

In line with previous observations [34], 6% of the multi-modal input turns proved to be concurrent, i.e., speech and gesture were synchronised, but semantically unrelated. For example, one user said "Denmark" to answer the system's question about the user's country of origin while gesturing on the picture of the Coliseum. Another user said "Where do you live?" while gesturing on the feather pen on the desk.

The evaluation of the GR, GI and IF modules can be summarised as follows:

- GR failures represent 12.8% of gestural inputs, but had no impact on interaction success.
- Failures in processing gesture-only input for *referenceable* objects involved the GI module in only 4% of the cases.
- Fusion failures occurred for 40% of the multi-modal behaviours. Three-fourth of these cases correspond to missing fusions and 1/4 to irrelevant fusions.

Thus, our comparative analysis of the video and log files shows that the gestures done on non-referenceable objects and the gestures done while the character was speaking or preparing to speak, had a quite negative impact on gesture interpretation. This is true both for the processing of gesture-only and multi-modal behaviours. Both might be due to the graphical affordance of referenceable objects and the lack of visibility of the non-verbal cues shown by the character. Indeed, graphical affordance could be improved in our system so that

(1) the users can visually detect the objects the character can speak of, e.g., these referenceable objects could be permanently highlighted, (2) the users understand that the character is willing to take or to keep the turn, e.g., the camera could be directed towards the character's face in such cases, thus enhancing the visibility of the non-verbal cues for turn-taking management. Our analysis also reveals how the dimensions of fusion were used by the user and processed by our system. We observed that the proper management of temporal information, such as the reception of a start of speech message at the right time has a huge impact on IF success. Regarding the semantic dimension, users only rarely did multi-object selection with a single gesture or made implicit spoken references to objects.

4.3. Interviews

Fig. 7 presents a summary of the users' answers in the post-test interviews. For each interview question, each user's answer was scored independently by two scorers on a 3-point scale from (1) positive with minor or no qualifications, over (2) positive with qualifications, to (3) negative/with substantial qualifications [35].

Six questions (Q(n)s) in the user interviews address gesture-related issues. On the question (Q3) *if Andersen was aware what the user pointed to,* most users were quite positive although some pointed out that Andersen ignored their gestures in some cases. This was expected due to the large number of non-referenceable objects in Andersen's study and is confirmed by the analysis in Section 4.2. The kids were almost unanimously positive in their comments on Q4, *how it was to use the touch screen,* which they found easy and fun. Like in the first prototype user interviews [2], the children were divided in their opinions on Q5 as to *whether they would like to do more with gesture.* Half of the users were happy with the 2D gesture affordances while the other half wished to be able to gesture towards more objects in Andersen's study. On the question (Q6) *whether they talked while pointing*, only a couple of users said that they never tried to talk and point at the same time. We will return to this point below. Finally, on the question (Q14) if the

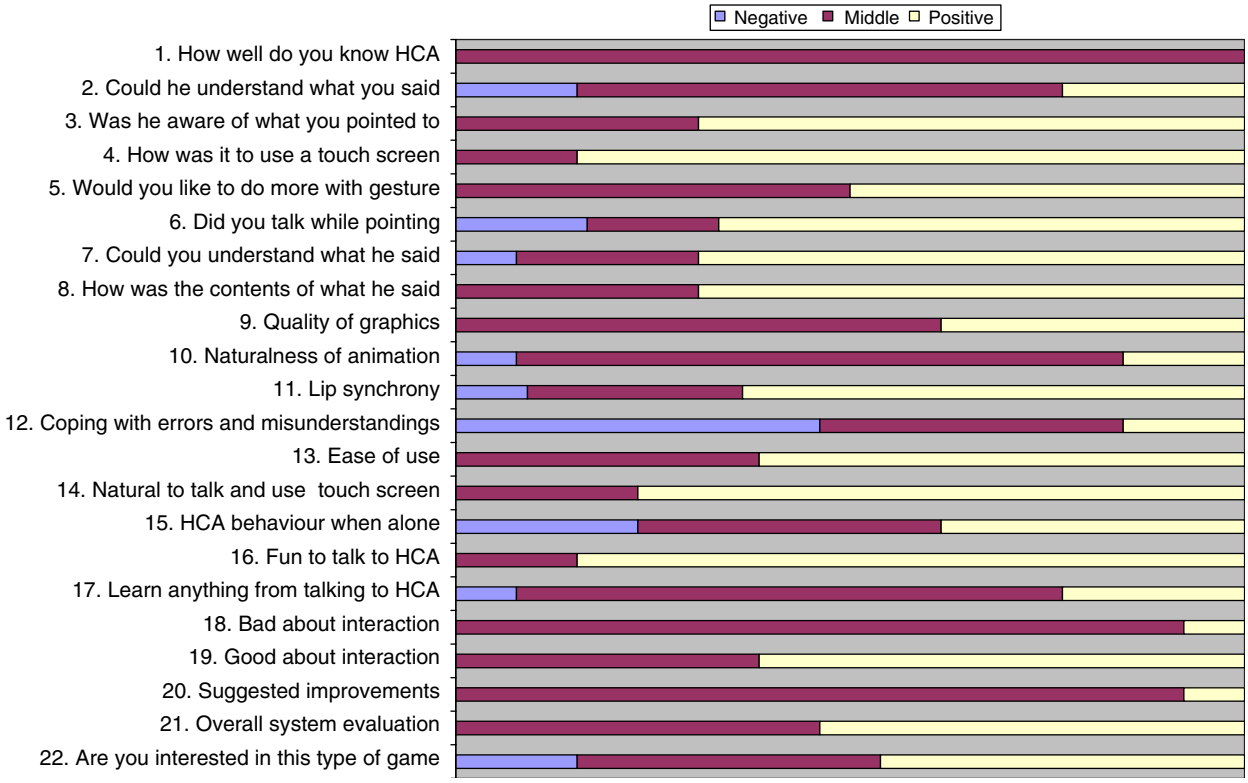

Fig. 7. Summary of interview results.

users *felt it to be natural to talk and use the touch screen,* the large majority of users were again quite positive.

In summary, the Danish users of the second Andersen prototype were almost unanimously happy about the available modality/device input combinations, i.e., pointing gesture input via touch screen and speech input via microphone headset (Q4, Q14). Andersen sometimes ignored the users' pointing gestures (Q3), which perhaps partly explains why half of the users wished to be able to elicit more stories from Andersen through gesture input (Q5). Finally, the majority of users claimed that they, at least sometimes, talked while pointing (Q6).

Globally, users were happy with gestural and multi-modal input and many wished to do more with gestures, which is congruent with previous observation that gesture is a key modality for young users to have fun and take initiative in the interaction [36].

4.4. Follow-up experiment with native English speakers

Following the second prototype user test, described above, with Danish children having English as their second language, we did a small user test with four children, two girls and two boys, 11–13 years old, all of whom had English as their first language. The primary purpose of the test was to explore the effects of (i) users' first language and (ii) the amount of instruction received on how to speak to the system. Thus, the English children were provided with extensive instructions on how to speak to the system during the first test condition, whereupon they carried out the second test condition in the same way as the Danish kids did, cf. Section 4.1. In what follows, we focus on a single finding in the test related to the Danish kids' response to Q6, i.e., that they often talked while gesturing.

To compare the Danish children with the English children, we randomly sampled four Danish children from the Danish user population, two girls and two boys. We then looked at the transcriptions from the directly comparable 2nd-condition trials in which all children were invited to address, at their leisure, topics from a list of 11 topics in conversation with Andersen. Table 7 shows what we found on the use of combined speech and gesture input in the two test groups.

Table 7
Combined speech and gesture input in two user groups

	Danish children	English children
No. of input turns	201	267
No. of speech–gesture turns	0	30
No. of speech–gesture turns per user	0-0-0-0	12-2-4-12
No. of gesture-only turns	15	4

Table 7 shows that the randomly sampled Danish users did *not* speak while gesturing at all. This is in sharp contrast to Danish group's response to (Q6) *whether they talked while pointing.* Even if, by (unlikely) chance, the sampled Danish group includes the two Danish users who admittedly never tried to talk and point at the same time, Table 7 includes four users who did not do that in the 2nd test condition. They might, of course, have done so in the first test condition. Whatever the explanation might be, this contrasts markedly with the English users, all of whom spoke when they gestured except in 12% of the turns in which they used gesture input. When the Danish kids in the sampled group used gesture, they never spoke at the same time.

The hypothesis arising from Table 7 is that there are significant behavioural differences between children having English as their first language and children having English as their second language, in the way they use the speech and gesture input affordances available. In order to obtain information on objects that can be indicated through gesture, the former naturally speak while gesturing whereas the latter tend to choose gesture input-only. The explanation for this hypothesis probably is that the opportunity to complete a conversation act without speaking a foreign language tends to be favoured whereas, for users speaking their mother tongue, it is more natural to speak and gesture at the same time. It should be noted here that the English users were very young, which speaks against attributing their more frequent use of multi-modal input to speaker maturity. This finding, hypothetical as it remains due to the small user populations involved, must be kept in mind when interpreting the results presented in this paper, most of which have been gathered with users having English as their second language.

5. Discussion

In this paper, we have presented early results on how 10–18 years old Danish children having English as their second language use speech and 2D gesture to express their communicative intentions in conversation with a famous 3D animated character from the past. In a small control study with 11–13 years old children having English as their first language, we found that the pattern of multi-modal interactive input apparent in the Danish kids might be significantly different in the English-speaking children. In essence, the English-speaking kids practice what the Danish children preach, lending strong joint support for the conclusion that the multi-modal input combination of speech and touch screen-enabled conversational input is a highly natural input combination for conveying users' communicative intentions to embodied conversational characters.

From a technical point of view, the work reported shows, first of all, that we are only at the very beginning of addressing the enormous challenges facing developers of natural interactive systems capable of understanding combined speech and 2D gesture input. In the following, we describe some of those challenges viewed from the standpoint of having completed and tested the 2nd Andersen system prototype.

5.1. Mouse vs. touch screen gesture input

It seems clear that gesture input via the touch screen device is far more natural for conversational purposes than gesture input via the mouse or similar devices, such as controllers. The mouse (controller) is a haptic input device, which a large user population is used to employ for, among other things, purposes of fast haptic control of computer game characters and other computer game entities. However, these input devices are far from being natural in the context of natural interactive *conversation.* When offered these devices, as we observed in the Prototype-1 user tests [30], the users tend to "click like crazy", following their—natural or trained—tendency to gesture around in the graphical output space without considering the conversational context. Conversely, when offered the more natural option of gesturing via the touch screen in a speech-gesture conversational input environment, no user seems to be missing the fast interaction afforded by the mouse (controller). On the contrary, given the interactive environment just described, users seem perfectly happy with gesturing via the touch screen, thereby emulating quite closely their real-life-familiar 3D pointing gestures, cf. Fig. 7, Question 4.

5.2. Referential disambiguation through gesture

While the Danish users clearly seem to have understood that they could achieve unambiguous reference to objects without having to speak, they also understood that spoken deictics require gesture for referential disambiguation. Confirming the users' claims about the intuitive naturalness of using touch screen-mediated 2D gesture, the children seem to be keenly aware of the need to point while referring in speech to the object pointed towards.

Another important point is that the users' coordinated spoken references to pointed-to objects were generally deictic in nature, making them amenable to handling by the IF component we had designed. Thus, in the large fraction of the 67 coordinated speech-gesture inputs in which the speech part actually did refer to the object(s) pointed towards, only one did not include deictics, i.e., "Would you please tell me about the watch".

5.3. Deictics fusion is only the tip of the iceberg

Essentially, the IF approach adopted for the Andersen system aims at semantic fusion of singular vs. plural spoken deictics with the number of named objects identified through gesture interpretation. IF also manages implicit or explicit references to concepts related to (system-internally) named objects in Andersen's study. For instance, "Do you like travelling" would be merged with a gesture on one particular object, i.e., Andersen's travel bag. What we found was that most users employed spoken deictics, i.e., pure demonstratives, such as 'this' in "What is this?" and only rarely used more explicit referential phrases, such as noun phrases.

However, even this simple fusion domain is subject to the fundamental ambiguity between, on the one hand, how many physical objects the user intends to refer to and, on the other, how many *within-object* entities the user intends to refer to, such as several objects depicted in a single picture. To resolve this ambiguity, the system would need knowledge about the inherent structure and contents of objects, such as pictures. Moreover, spoken

deictics do not necessarily refer to gestured-towards objects. It is perfectly normal for spoken deictics to anaphorically refer to the spoken discourse context itself, as in "Are these your favourite fairytales?" Given the fact that users sometimes perform mutually independent (or concurrent) conversation acts through speech and gesture, respectively, the system would need quite sophisticated meta-communication defences to pick up the fact that the user is not performing a single to-be-fusioned conversation act but, rather, two quite independent conversation acts. Finally, requiring the system to be able to manage, and hence to have knowledge about, the internal structure and contents of objects, such as pictures, is a demanding proposition. In the foreseeable future, we would only expect highly domain-specific applications to be able to handle this problem, such as museum applications for users to inquire about details in museum exhibit paintings.

5.4. Other chunks of the iceberg

As we saw in Section 4, users may, in principle, point to anything in Andersen's study and speak at the same time. Furthermore, what they may relevantly say when gesturing is open-ended, including, for instance, the volunteered conversation act ⟨pointing to a chair⟩ "My grandfather has a chair like that". This conversation act is relevant simply because Andersen's study is one of the system's domains of conversation. Users may also explore relationships among objects, requiring the character to have a model of these, as in ⟨pointing to picture of Coliseum⟩ "Do you have other pictures from your travels?"

We do not believe that the current Andersen system architecture (Fig. 3) is the best solution for handling the just illustrated, full-scale speech-gesture IF for domain-oriented conversation. At the very least, it seems, NLU must be made aware that the currently processed spoken input is being accompanied by gesture input. Otherwise, the complexity to be handled by IF is likely to become monstrous. An even better solution may be to process speech and gesture input together, removing the need for a subsequent late semantic IF component. As regards conversation management (in the character module) and response generation, on the other hand, we see no evident obstacles for the current architectures to process far more complex IF than what is currently being processed by the Andersen system.

In conjunction with Andersen's injunctions to do so, the design of Andersen's study did lead the users to gesture at the pictures on the walls. Inevitably, however, these factors also made the users try to find out which objects Andersen could actually tell stories about. In the first Andersen prototype, we had an additional class of "anonymous objects" which were referenceable, but which, when gestured upon, made Andersen say that he did not know much about them at present. In the second prototype, we dropped this class because it was felt that Andersen's response was not particularly informative or interesting, and tended to be tedious when frequently repeated. Since, for Prototype-2, we did not increase the number of objects which Andersen had stories to tell about, the consequence was an increase in the number of failures in gesture interpretation and IF since the users continued to gesture at objects which were presented graphically, but which the system did not know about (i.e., the non-referenceable objects). There is no easy solution to this problem. One solution is to increase the number of objects which Andersen can tell stories about until that number converges with the objects which the majority of users want to know about. Another solution is to make Andersen know about all objects in his study, including the ceiling and the carpet. A third, more heavy-handed and less natural, solution might be to have specific rendering for the objects the user can gesture at to get Andersen to tell about them, such as by using some form of permanent highlighting.

The user did not use the cross shape in their gestures. This might be due to the fact that this gesture shape is not that appropriate for the tactile screen.

Selection of several objects in a single gesture, using, e.g., encirclement or a connecting line, never occurs in our data. Nor does the data show a single case of plural spoken deictics, such as "these books". This may be due in part to the fact that the placement of the individual objects on the walls of Andersen's study did not facilitate the making of connections between them, and partly to the relative scarcity of our data. Arguably, sooner or later, a user might say, ⟨pointing to the books on the bookshelf⟩ e.g., "Tell me about these books". We did not observe perceptual grouping behaviours, e.g., using a deictic plural in speech, such as "these pictures", and selecting a single picture in a group of pictures with a pointing gesture.

This might be due to several reasons. It was not demonstrated in the simple multi-modal example the users were shown at the start of the test. Another reason might be the current layout of the graphical objects and the richness of their perceptual properties (e.g., the pictures) as compared to the 2D geometric shapes investigated in [32].

As we explained in the analysis of the users' multi-modal behaviour, users nearly always used spoken deictics (pure demonstratives) rather than actually naming the objects referred to, probably because this was included in the short demonstration they had prior to the experiment and because the recognition of deictics happened to work quite well. They nevertheless also used a variety of references that were not demonstrated (e.g., "who is this woman?"), showing that they were able to generalise to other kinds of references. This nevertheless raises the issue of natural vs. trained multi-modality [37]. On the one hand, full natural multi-modality (e.g., not showing any gesture or multi-modal examples to the users prior to testing) will probably lead to an even smaller proportion of multi-modal behaviours than the one we observed. On the other hand, trained multi-modality might generate a larger variety of examples, such as multiple-object gestures and implicit spoken references without any deictics. We believe that the approach we selected, i.e., that of demonstrating a single example of a multi-modal input combination, is a reasonable trade-off between these two extremes.

It follows that there are a serious number of challenges ahead in order to be able to handle natural interactive speech-gesture conversation, including issues arising from the Andersen system, such as:

1. the plural deictics/one object problem (the user refers to several items in a single picture);
2. demonstratives may refer to spoken discourse as well as to the visual environment;
3. addressing object details: a very demanding proposition for developers;
4. addressing—potentially several—objects by a (user-) stated criterion, such as "Can you show me all the pictures from your fairytales?"
5. users may point at anything visible (and possibly ask as well);
6. users may meaningfully ask about, or comment on, objects without pointing, as in "Who painted the portrait of Jenny Lind?"
7. using visible objects as illustrations in spoken discourse.

However, as regards the children who participated in the Prototype-2 user test, only Point 5 posed a significant problem, whereas Points 1 and 3 posed minor problems. Points 4, 6 and 7 never occurs in the data whereas Point 2 occurs a few times.

6. Conclusions

In this paper, we have described the modules that we have developed for processing gesture and multi-modal input in the Andersen system, as well as their evaluation with two different groups of young users. We have identified the causes of the most frequent module failures, i.e., end of speech management in the speech recogniser, gestures on non-referenceable objects, and input gesturing while the character is preparing to speak. We have suggested possible improvements for removing these errors, such as improvement of graphical and non-verbal affordance, and proper management of end of speech messages by the speech recogniser.

The Andersen project described in this paper has provided data on how children gesture and combine their gesture with speech when conversing with a 3D character. Below, we revisit the issues that were raised in the introduction.

How do children combine speech and gesture? They do so more or less like adults do but (i) probably in a slightly simpler fashion and (ii) only if they are first-language speakers of the language used for interaction with the ECA.

Would children avoid using combined speech and gesture if they can convey their communicative intention in a single modality? No, not if they are first-language speakers of the language used in the interaction; but yes, if the language of interaction is their second language.

Is their behaviour dependent upon whether they use their mother tongue or a second language? This seems likely to be the case, but we need more data analysis for confirmation.

To what extent would the system have to check for semantic consistency between the speech and the perceptual features of the object(s) gestured at? We observed that the recognition and understanding of spoken deictics was quite robust in the system and that spoken deictics were nearly always used in multi-modal input. We also observed behaviour in

which there was semantic inconsistency between the speech and the perceptual features of the gestured object. One user would ask "Who is this woman?" when pointing to the picture of a man. This man is wearing old-fashioned clothes and the picture, which is in the corner of the room, might be less visible than the other pictures. Another user would say, "What is this?" when pointing to a picture showing the picture of Andersen's mother. We might have expected "Who is this?" Finally, the difficulties of speech recognition observed show that it was better for the system to primarily trust the gesture modality as it appeared, and was expected, to be more robust than the speech. Since this paper focused on gesture and combined speech-gesture in the Prototype-2 user tests, we have not discussed the speech processing findings made in those tests. Suffice it here to say that the percentage of perfect speech recognition was 23% for the Danish users and 33% for the English users, whereas the percentages for perfect gesture recognition and interpretation were in the range of +90% for both user groups. The system's 2000 words speech recogniser vocabulary was adequate for recognising and understanding the spoken parts of the users' multi-modal input despite the fact that the vocabulary had been developed on the basis of spoken-input-only corpora.

How do we evaluate the quality of such systems? In this paper, we have used standard evaluation methodologies, technical as well as usability-related, for assessing the quality of the design solutions adopted for gesture and combined speech-gesture input processing. The solutions themselves represent relatively complex trade-offs within the, still partially uncharted, design space for multi-modal speech/gesture input systems.

Some more specific evaluation methodologies have also been considered in the literature. For example, in their book dedicated to the evaluation of ECAs, [45] point out the difference between micro-level evaluation focused on a single feature of the ECA and macro-level evaluation focused on the global contribution of the ECA to an application. In the same book, [38] provide a taxonomy of macro-level dimensions to evaluate in an ECA, such as believability or sociability, with corresponding evaluation criteria. Another evaluation issue concerns the target users of the Andersen system, i.e. children and teenagers, who may require some specific methods to optimise the data collection. In this respect, [39] recommend methods, such as thinking aloud, peer tutoring or user diaries in order to access children's mental model and unbiased comments on a system. The authors also point out the inadequacy of using some methods with children, such as the use of focus groups. Finally, the context of a game application raises additional evaluation issues in the Andersen project, because a game has to be usable and challenging at the same time in order to be entertaining [40,41]. Computer games can be evaluated by complementary means, such as classical usability methods, psycho-physiological measures and behavioural analysis [42–44]. However, among all these methodologies—for evaluation of ECAs, doing tests with children, and evaluating computer games—none especially focus on investigating multi-modal input. Therefore, we chose to rely on classical methods for this particular topic, and we might draw on those specific methods for evaluating other dimensions of the Andersen system, e.g., Andersen's believability and entertainment qualities.

What do the users think of ECA systems affording speech and gesture input? They clearly like to use the touch screen and they very much appreciate the idea of combined speech-gesture input even if they do not massively practice combined speech-gesture input when the language of interaction is not their first language. Speech and gesture input is, indeed, a "natural multi-modal compound" for ECA systems.

How to manage temporal relations between speech input, gesture input and multi-modal output? We have proposed algorithms for managing the temporal dimension and provided an illustration of the multiple considerations involved when the system is large and complex. According to our evaluation, as reported above, the algorithms proved suitable for the management of the users' behaviour.

The data we have collected clearly needs to be complemented by data obtained with behaviours in other multi-modal conversational contexts, possibly more complex regarding graphical affordance for multi-modal behaviour, such as many different types of graphical objects, complex occlusion patterns, etc. This might elicit more ambiguous gesture semantics requiring the management of gesture confidence scores, speech confidence scores being notoriously unreliable for many important purposes.

In the current state of the art in the field of embodied conversational agents, Andersen is

probably one-of-a-kind. We know of no other running system, which integrates solutions to the challenges listed in Section 1.1. There is a sense in which the Andersen system is simply a computer game with spontaneous spoken interaction between the user and the character. This field of interactive spoken computer games was close to non-existent when the NICE project began. Spoken *output* in computer games was commonplace when the project began, however. Today, several computer games offer spoken input command words, which make a game character perform some action. So far, these products do not seem terribly popular with the games reviewers, probably because they typically assume that the game player is able to learn, sometimes quite large, numbers of spoken commands, and because their speech recognition and understanding is too fragile as well. We are not aware of any interactive spoken computer game products in the market. This is hardly surprising. Viewed from the perspective of the Andersen system, it may be too early to offer customers interactive spoken computer games in the standard sense of the term "computer game", knowing that a computer game is being used, on average, for 30–50 h of game-playing. By contrast, the Andersen system addresses the more modest challenge of providing edutaining conversation with a new user every 5–20 min.

Acknowledgements

We gratefully acknowledge the support for the NICE project by the European Commission's Human Language Technologies Programme, Grant IST-2001-35293. We would also like to thank all participants in the NICE project for the three productive years of collaboration that led to the running system prototypes presented in this paper.

References

[1] R.A. Bolt, "Put-that-there": voice and gesture at the graphics interface, Seventh Annual International Conference on Computer Graphics and Interactive Techniques, ACM, Seattle, Washington, US, 1980, pp. 262–270.

[2] N.O. Bernsen, L. Dybkjær, Evaluation of Spoken Multimodal Conversation. Sixth International Conference on Multimodal Interaction (ICMI'2004), Association for Computing Machinery (ACM), New York, 2004, pp. 38–45.

[3] S.L. Oviatt, Multimodal interfaces. Human–computer interaction handbook: fundamentals, in: J. Jacko, A. Sears (Eds.), Evolving Technologies and Emerging Applications, vol. 14, Lawrence Erlbaum Assoc., Mahwah, NJ, 2003, pp. 286–304.

[4] R. Sharma, M. Yeasin, N. Krahnstoever, I. Rauschert, G. Cai, I. Brewer, A. MacEachren, K. Sengupta, Speech–gesture driven multimodal interfaces for crisis management, Proc. IEEE VR2004 91 (9) (2003) 1327–1354 http://spatial.ist.psu.edu/cai/2003-Gesture-speeech-interfacesfor%20crisis-management.pdf.

[5] R. Catizone, A. Setzer, Y. Wilks, Multimodal Dialogue Management in the COMIC Project. EACL 2003 Workshop on Dialogue Systems: Interaction, Adaptation, and Styles of Management, 2003, http://www.hcrc.ed.ac.uk/comic/documents/publications/eaclCOMICFinal.pdf.

[6] M. Johnston, Unification-based multimodal parsing, 17th International Joint Conference of the Association for Computational Linguistics. Montreal, Canada. Association for Computational Linguistics, Morristown, NJ, USA, 1998.

[7] M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman, I. Smith, Unification-based Multimodal Integration, ACL'97, 1997.

[8] L. Almeida, I. Amdal, N. Beires, M. Boualem, L. Boves, E. Os, P. Filoche, R. Gomes, J.E. Knudsen, K. Kvale, J. Rugelbak, C. Tallec, N. Warakagoda, The MUST Guide to Paris; Implementation and expert evaluation of a multimodal tourist guide to Paris. Multi-Modal Dialogue in Mobile Environments, ISCA Tutorial and Research Workshop (IDS'2002), Kloster Irsee, Germany, June 17–19 http://www.isca-speech.org/archive/ids_02, 2002.

[9] M. Johnston, S. Bangalore, Multimodal Applications from Mobile to Kiosk. W3C Workshop on Multimodal Interaction, Sophia Antipolis, France, 19–20 July 2004, 2004 http://www.w3.org/2004/02/mmi-workshop/papers

[10] S. Oviatt, Multimodal interactive maps: designing for human performance, Hum. Comput. Interact. 12 (1997) 93–129.

[11] A.D. Milota, Modality Fusion For Graphic Design Applications, ICMI'2004, 2004.

[12] P. Gieselmann, M. Denecke, Towards multimodal interaction with an intelligent room. Eighth European Conference On Speech Communication and Technology (Eurospeech'2003), Geneva, Switzerland, September 1–4, 2003, http://isl.ira.uka.de/fame/publications/FAME-A-WP10-007.pdf.

[13] J. Juster, D. Roy, Elvis: situated speech and gesture understanding for a robotic chandelier. Sixth International Conference on Multimodal Interfaces (ICMI'2004), October 13–15, State College, Pennsylvania, USA, ACM, New York, 2004, pp. 90–96.

[14] S.L. Oviatt, R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson, L. Carmichael, Toward a theory of organized multimodal integration patterns during human–computer interaction, in: International Conference on Multimodal Interfaces (ICMI'2003), ACM Press, Vancouver, BC, 2003, pp. 44–51 http://www.cse.ogi.edu/CHCC/Publications/toward_theory_organized_multimodal_integration_oviatt.pdf.

[15] W.C. Avaya, D. Dahl, M. Johnston, R. Pieraccini, D. Ragget, EMMA: Extensible MultiModal Annotation markup language. W3C Working Draft 14 December 2004, W3C. http://www.w3.org/TR/emma/

[16] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, S. Feiner, Mutual disambiguation of 3D

multimodal interaction in augmented and virtual reality, in: Fifth International Conference on Multimodal Interfaces (ICMI'03), ACM Press, Vancouver, British Columbia, Canada, 2003, pp. 12–19 http://www1.cs.columbia.edu/~aolwal/projects/maven/maven.pdf.

[17] J. Cassell, J. Sullivan, S. Prevost, E. Churchill, Embodied Conversational Agents, MIT Press, Cambridge, MA, 2000, 0-262-03278-3.

[18] W.L. Johnson, J.W. Rickel, J.C. Lester, Animated pedagogical agents: face-to-face interaction in interactive learning environments, Int. J. Artif. Intell. Educ. 11 (2000) 47–78 http://www.csc.ncsu.edu/eos/users/l/lester/www/imedia/apa-ijaied-2000.html.

[19] D. Traum, J. Rickel, Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds, First International Joint Conference on "Autonomous Agent and Multiagent Systems" (AAMAS'02), July 15–19, Bologna, Italy, ACM Press, New York, 2002, pp. 766–773.

[20] T. Sowa, S. Kopp, M.E. Latoschik, A Communicative Mediator in a Virtual Environment: Processing of Multimodal Input and Output, In: Proc. of the International Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy, 2001, pp. 71–74, http://www.techfak.uni-bielefeld.de/~skopp/download/CommunicativeMediator.pdf.

[21] D. Hofs, H.J.A. op den Akker, A. Nijholt, A generic architecture and dialogue model for multimodal interaction. 1st Nordic Symposium on Multimodal Communication, Copenhahen, Denmark, 25–26 September 2003, pp. 79–92.

[22] S. Narayanan, A. Potamianos, H. Wang, Multimodal systems for children: building a prototype, Sixth European Conference on Speech Communication and Technology (Eurospeech'99), Budapest, Hungary, September 5–9, 1999.

[23] D. Perzanowski, A.C. Schultz, W. Adams, E. Marsh, M. Bugajska, Building a multimodal human-robot interface, IEEE Intell. Syst. 16 (1) (2001) 16–21.

[24] H. Holzapfel, K. Nickel, R. Stiefelhagen, Implementation and Evaluation of a Constraint-Based Multimodal Fusion System for Speech and 3D Pointing Gestures. ICMI 2004, 2004, http://isl.ira.uka.de/fame/publications/FAME-A-WP10-028.pdf.

[25] S. Oviatt, C. Darves, et al., Toward Adaptive Conversational Interfaces: Modeling Speech Convergence with Animated Personas, ACM Transactions on Computer–Human Interaction (TOCHI) 11(3) (2004).

[26] S.L. Oviatt, B. Adams, Designing and evaluating conversational interfaces with animated characters, in: J. Cassell, J. Sullivan, S. Prevost, E. Churchill (Eds.), Embodied Conversational Agents, MIT Press, Cambridge, MA, 2000, pp. 319–345.

[27] K. Ryokai, C. Vaucelle, J. Cassell, Virtual peers as partners in storytelling and literacy learning, J. Comput. Assist. Learn. 19 (2003) 195–208.

[28] J. Read, S.MacFarlane, C. Casey, Oops! silly me! errors in a handwriting recognition-based text entry interface for children. Nordic Conference on Human-Computer Interaction (NordiCHI '02), 2002, pp. 35–40.

[29] N.O. Bernsen, M. Charfuelàn, A. Corradini, L. Dybkjær, T. Hansen, S. Kiilerich, M. Kolodnytsky, D. Kupkin, M. Mehta, First prototype of conversational H.C. Andersen. International Working Conference on Advanced Visual Interfaces (AVI'2004), Gallipoli, Italy, May 2004, ACM, New York, 2004, pp. 458–461.

[30] S. Buisine, J.-C. Martin, N.O. Bernsen, Children's Gesture and Speech in Conversation with 3D Characters. HCI International 2005, Las Vegas, USA, 22–27 July 2005.

[31] E. Lewin, "KTH Broker, 1997, http://www.speech.kth.se/broker/

[32] F. Landragin, N.Bellalem, L. Romary, Visual salience and perceptual grouping in multimodal interactivity. First International Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy, 2001, pp. 151–155, http://www.loria.fr/~landragi/publis/ipnmd.pdf.

[33] S. Oviatt, A. De Angeli, K. Kuhn, Integration and synchronization of input modes during multimodal human–computer interaction, in: Human Factors in Computing Systems (CHI'97), ACM Press, New York, 1997, pp. 415–422.

[34] S. Buisine, J.-C. Martin, Children's and Adults' Multimodal Interaction with 2D Conversational Agents. CHI'2005, Portland, Oregon, 2–7 April 2005.

[35] N.O. Bernsen, L. Dybkjær, User evaluation of Conversational Agent H. C. Andersen, Ninth European Conference on Speech Communication and Technology (Interspeech'2005), Lisboa, Portugal, 2005.

[36] S. Buisine, J.-C. Martin, Experimental evaluation of bi-directional multimodal interaction with conversational agents, Proceedings of the Ninth IFIP TC13 International Conference on Human–Computer Interaction (INTERACT'2003), Zürich, Switzerland, September 1–5, IOS Press, 2003, pp. 168–175, http://www.interact2003.org/.

[37] J. Rugelbak, K. Hamnes, Multimodal Interaction—Will Users Tap and Speak Simultaneously? Telektronikk, 2003, http://www.eurescom.de/~ftproot/web-deliverables/public/P1100-series/P1104/Multimodal_Interaction_118_124.pdf.

[38] K. Isbister, P. Doyle, The blind men and the elephant revisited, in: Z. Ruttkay, C. Pelachaud (Eds.), From Brows to Trust: Evaluating Embodied Conversational Agents, Kluwer Academic Publishers, Dordrecht, 2004, pp. 3–26.

[39] S. MacFarlane, J. Read, J. Höysniemi, P. Markopoulos, Evaluating interactive products for and with children. Tutorial Notes, Interact'2003 Conference, 2003.

[40] D. Johnson, J. Wiles, Effective affective user interface design in games, Ergonomics 46 (2003) 1332–1345.

[41] K. Keeker, R. Pagulayan, J. Sykes, N. Lazzaro, The untapped world of video games, CHI'2004, 2004, 1610–1611,

[42] S. Kaiser, T. Wehrle, S. Schmidt, Emotional episodes, facial expressions, and reported feelings in human-computer interactions, Proceedings of Conference of the International Society for Research on Emotions, 1998, pp. 82–86.

[43] N. Lazzaro, K. Keeker, What's my method? A game show on games, in: Proceedings of CHI'2004, 2004, pp. 1093–1094.

[44] R. Pagulayan, K. Keeker, D. Wixon, R.L. Romero, T. Fuller, User-centered design in games, in: J.A. Jacko, A. Sears (Eds.), The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications Archive, Lawrence Erlbaum Associates, Inc., Mahwah, 2002, pp. 883–906.

[45] Z. Ruttkay, C. Pelachaud, From Brows to Trust—Evaluating Embodied Conversational Agents, Kluwer, 1-4020-2729-X, 2004, http://wwwhome.cs.utwente.nl/~zsofi/KluwerBook.htm.
[46] B. Xiao, C. Girand, S.L. Oviatt, Multimodal Integration Patterns in Children, in: J. Hansen, B. Pellom (Ed.), Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP'2002), Denver, CO, Casual Prod. Ltd., Sept. 2002, pp. 629–632. Abstract, http://www.cse.ogi.edu/CHCC/Publications/multimodal_integration_patterns_in_children_xiao.pdf.